Latency in PCIe® Expansion Systems

Kevin Burt

Technical Marketing

Samtec, Inc

Disclaimer

<u>Presentation Disclaimer</u>: All opinions, judgments, recommendations, etc. that are presented herein are the opinions of the presenter of the material and do not necessarily reflect the opinions of the PCI-SIG®.

Acknowledgments

Hugo Kohmann **Dolphin Interconnect Solutions**

Preben N. Olsen **Dolphin Interconnect Solutions**

Agenda

PCIe® Expansion

Optical Interconnects Overview

Latency Components

- Physical layer
- Link Layer
- Transaction Layer
- Operating System

Measurement Results

- Copper vs. Optical
- Linux and Real Time Operating System

Summary

PCIe Extension Introduction

Introduction

PCIe is a low latency robust protocol

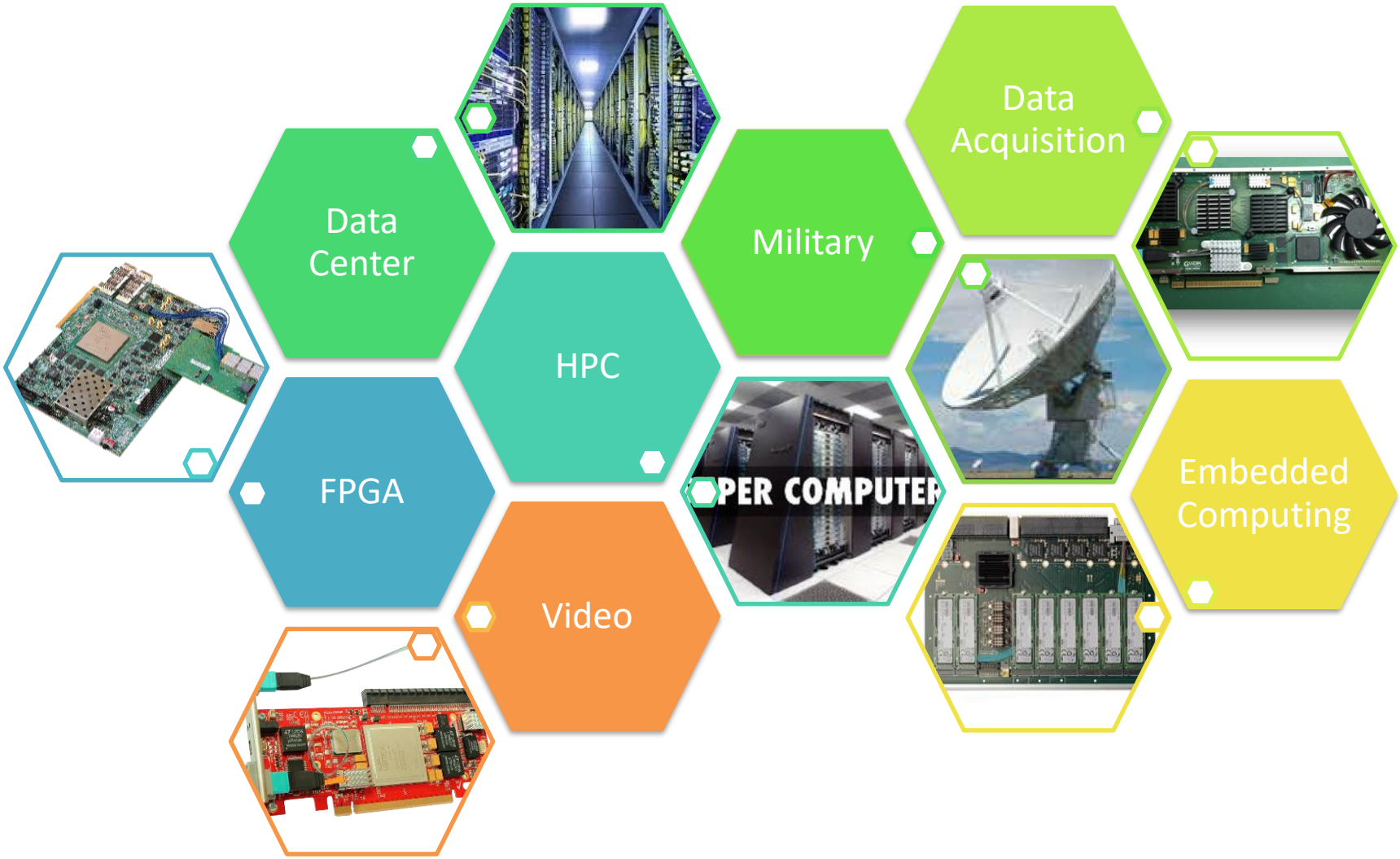
Traditionally, PCIe is used inside the box resulting in short channels.

PCIe extension lengthens these channels to provide the PCIe benefits to larger systems

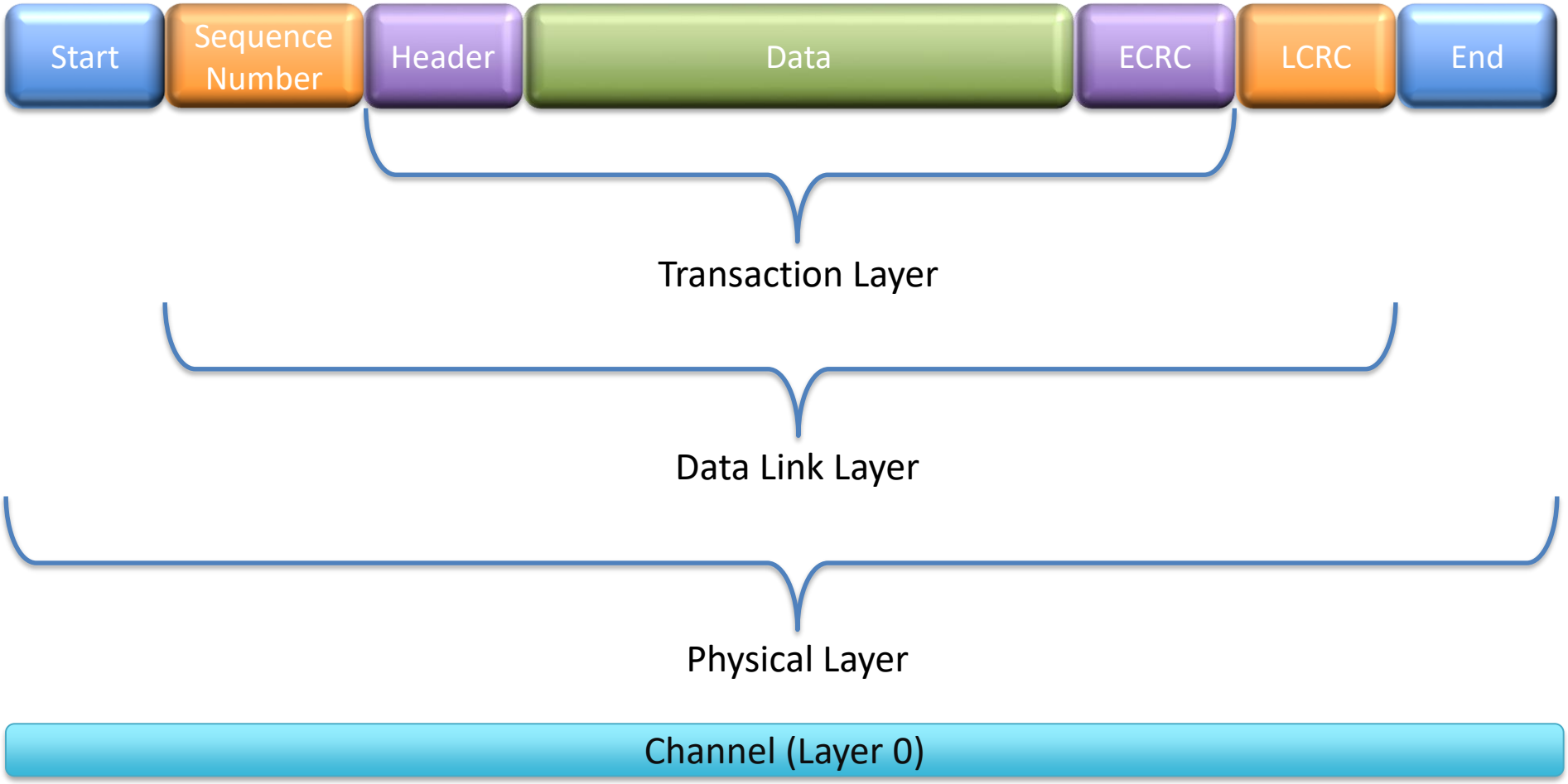
- PCIe over Copper ≤ 7 m
- PCIe over fiber > 100 m

PCIe protocol is designed to accommodate these longer lengths

Applications

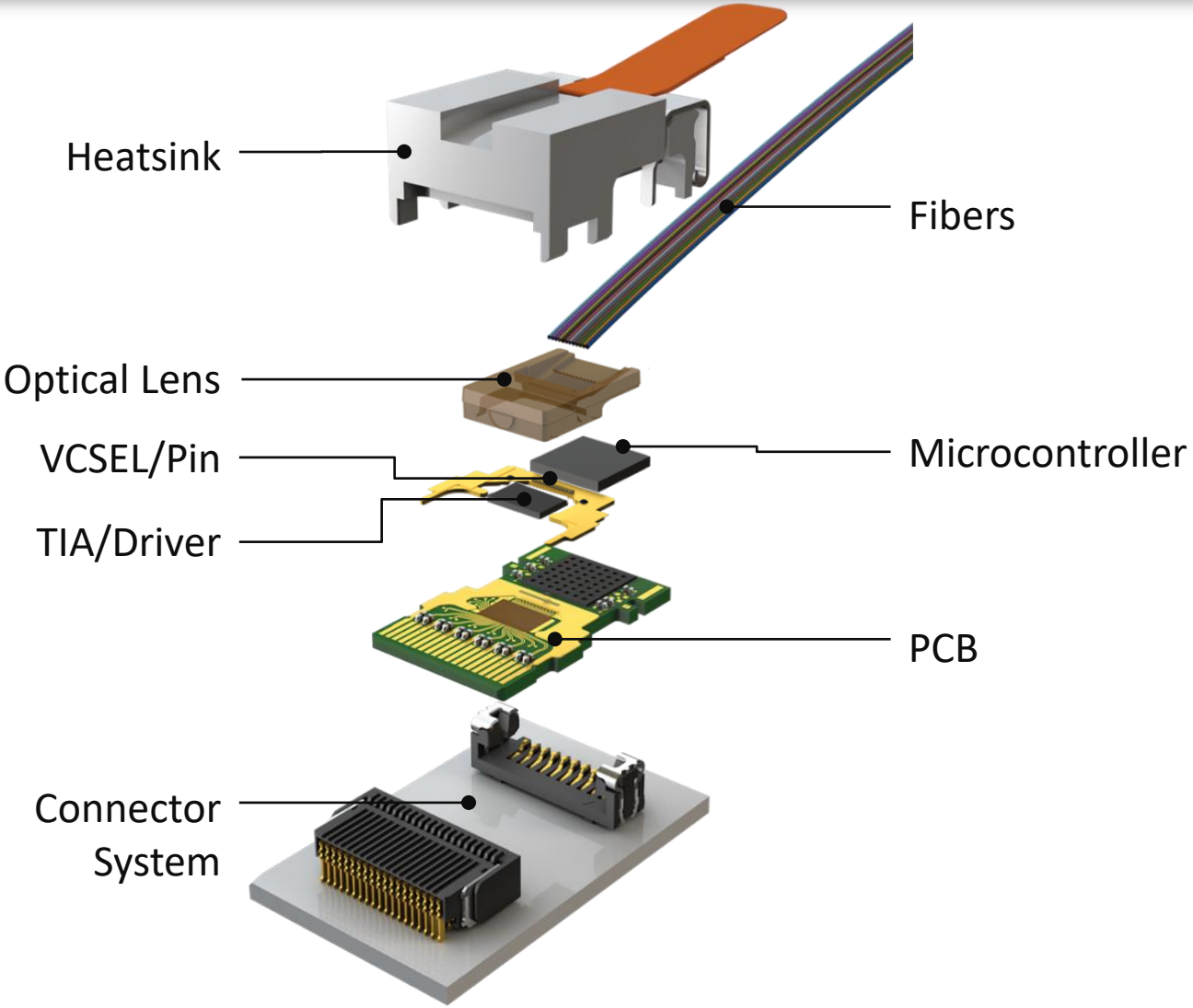


PCIe Layers

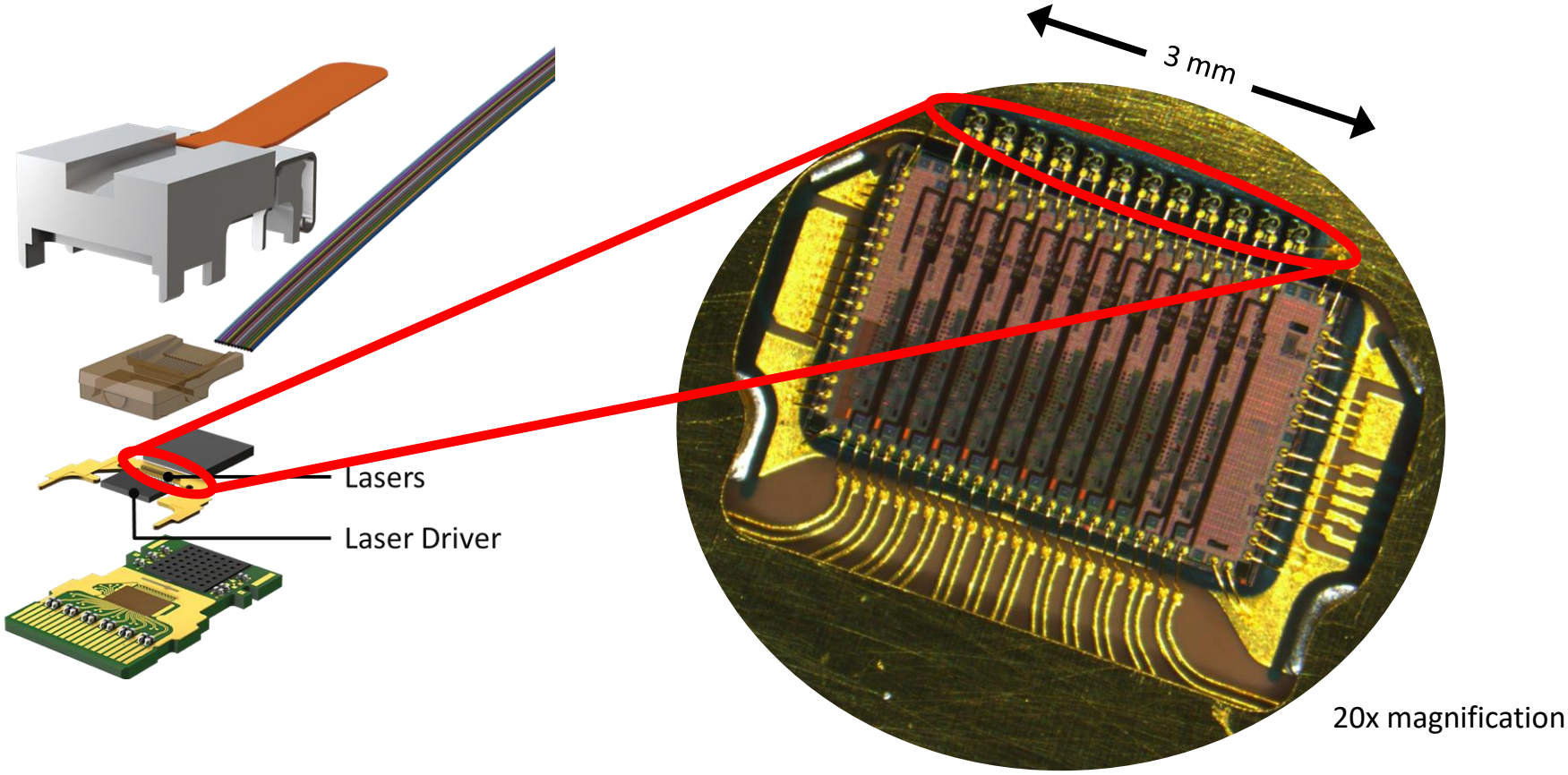


Optical Interconnects

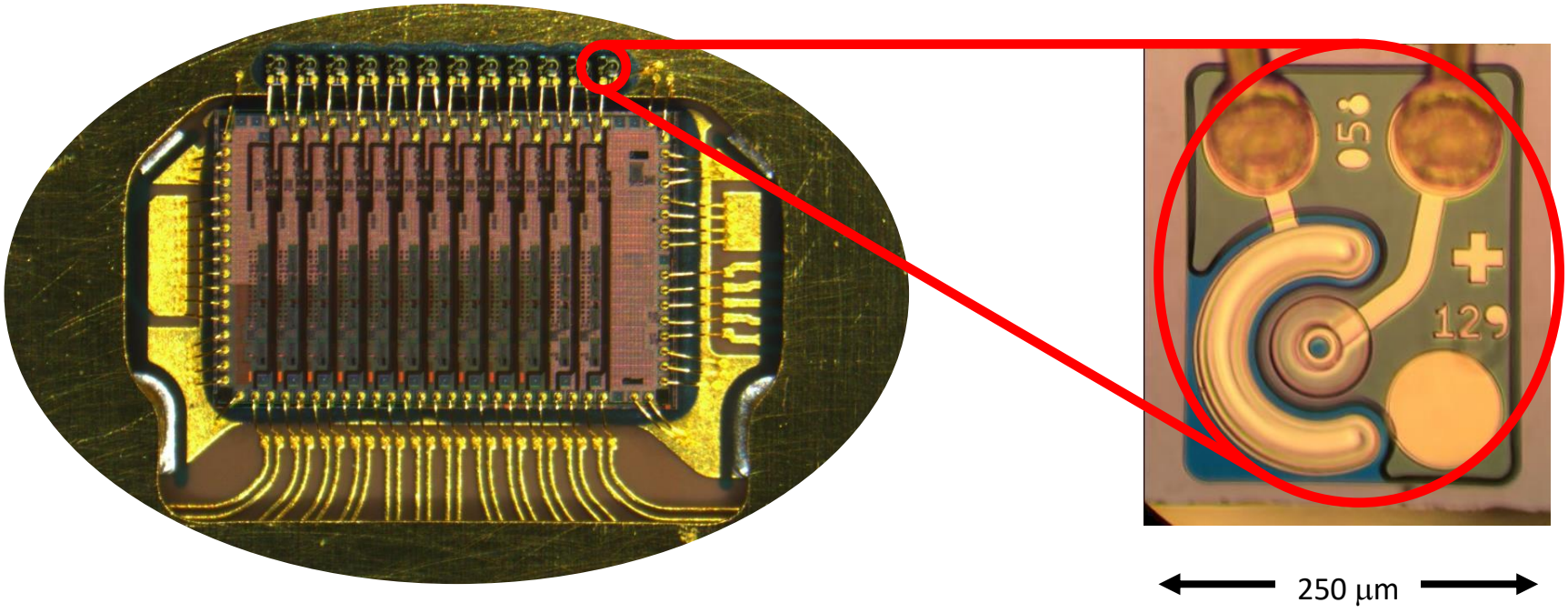
Anatomy of an Optical Interconnect



Laser Scale

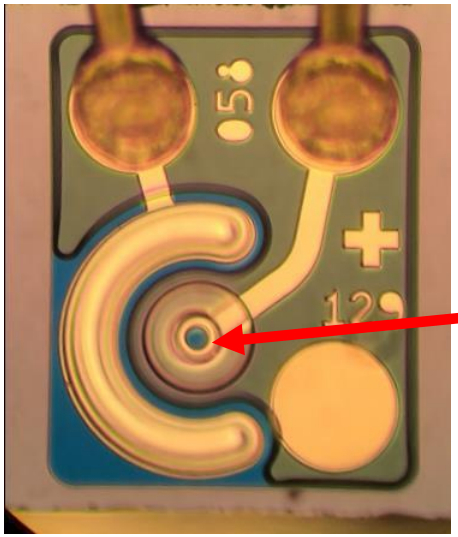


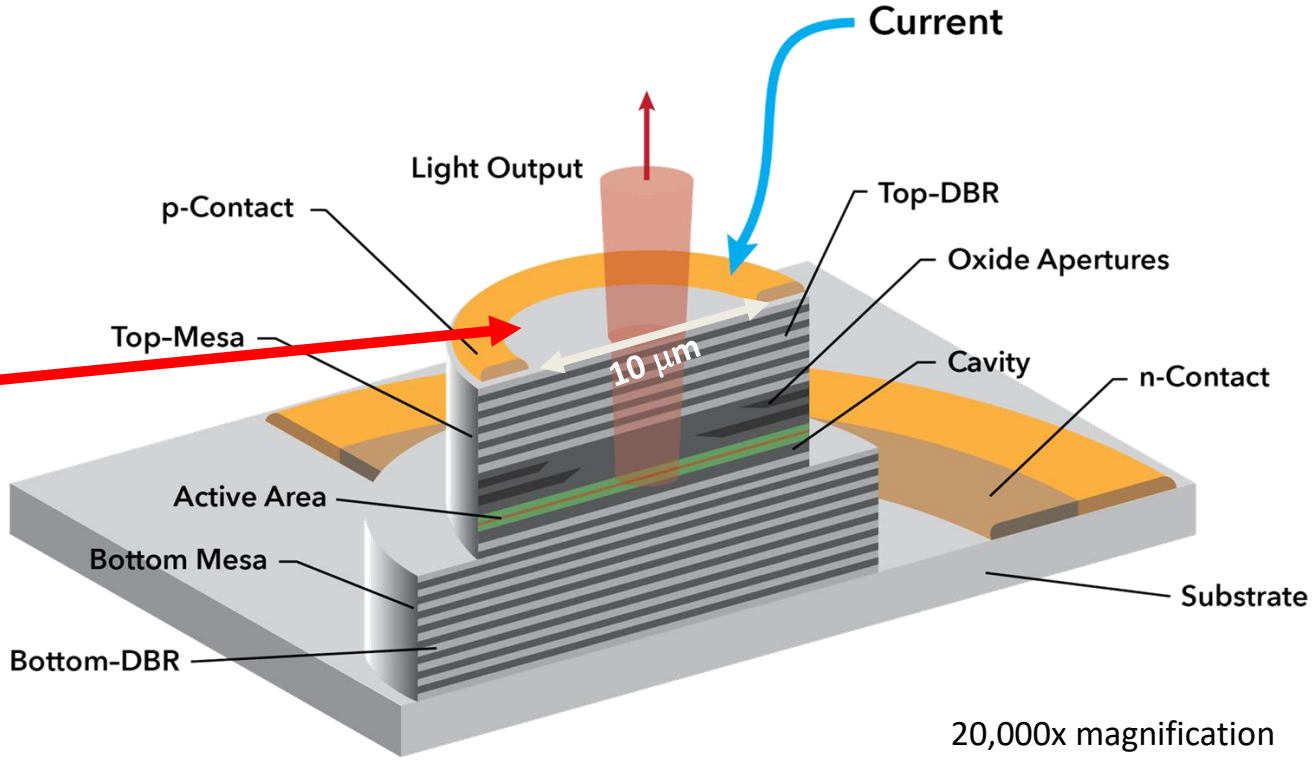
Laser Scale



200x magnification

Laser Scale

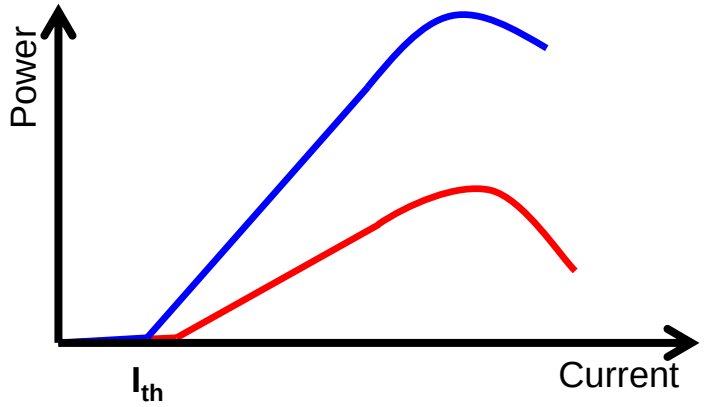




20,000x magnification

Laser Output Power vs. Input Current

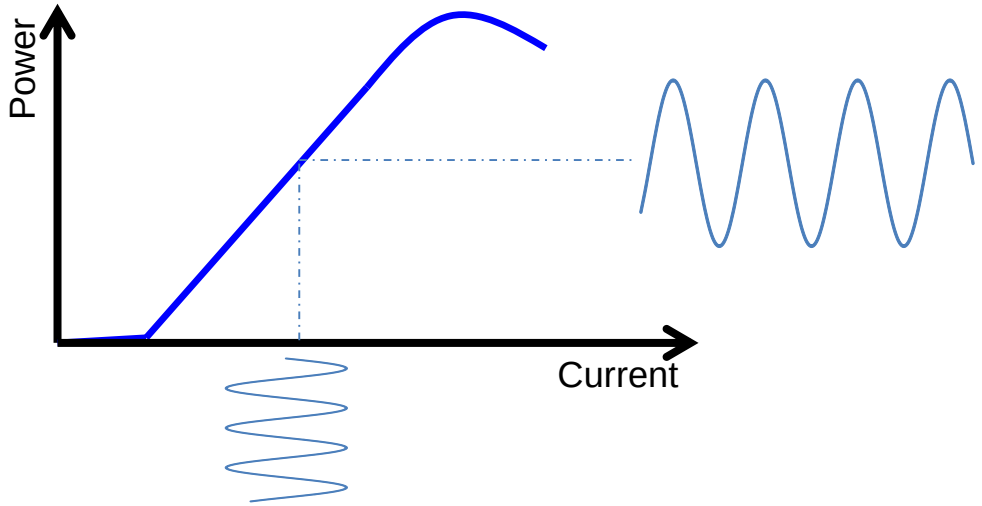
The average output power from a laser increases with the input current



- At low currents, the power is small
- Above the *threshold current,* I_{th}, the power increases linearly with current
- As the current increases, the temperature rises and the laser loses efficiency. The slope *rolls over* and the power drops
- At hot temperatures, the threshold current increases and the efficiency drops

Transmitter- Direct Modulation of the Laser

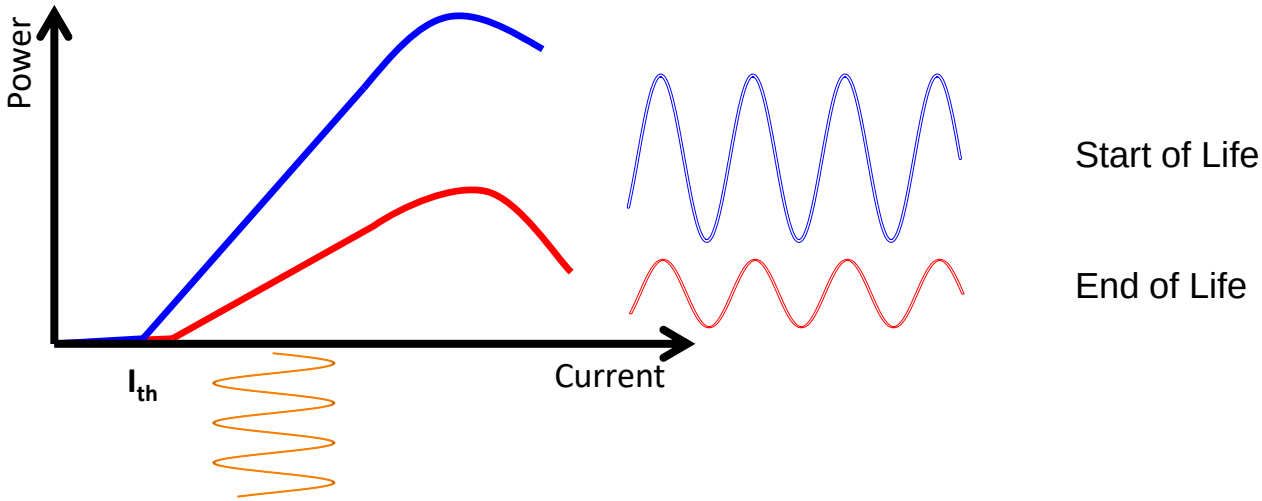
The average output power from a laser increases with the input current



- A bias current is used to operate laser safely above threshold and below rollover
- A modulation current is used to encode the data and provide the variation in the output power that carries the (digital) signal

Change with Time

Laser changes over life



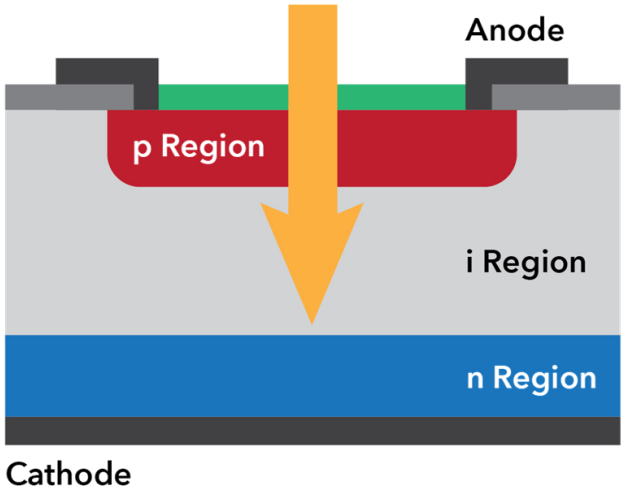
- Power decreases
- Peak to peak swing decreases

PIN

Converts light back to a (small) current

Design is a compromise between high speed performance and ability to convert light

- Bigger = better ability to capture and convert
- Smaller = less capacitance and higher bandwidth



Also adds noise:

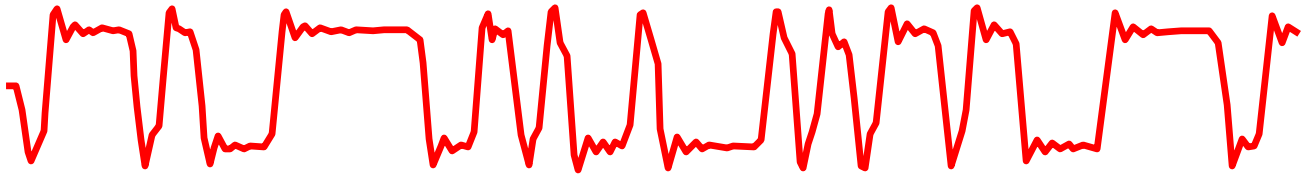
- Shot Noise, I_S
- Thermal noise, I_T
- Dark Current noise, I_D

$$Total\ Noise, I_{noise} = \sqrt{I_S^2 + I_T^2 + I_D^2}$$

Receiver Function

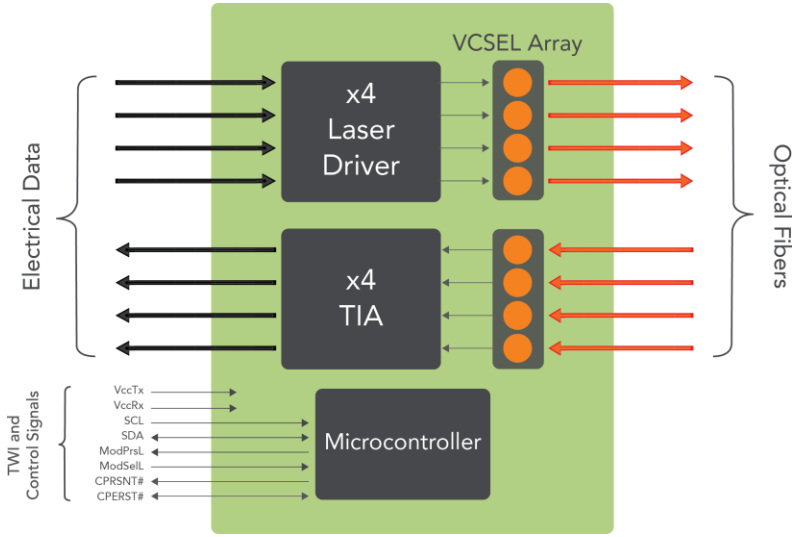
Digital Data is always 1s and 0s where each bit is the same length

Received Data is not that simple:

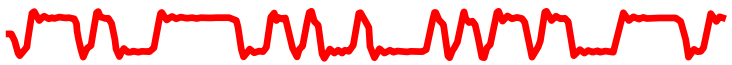


We need to recover the original signal:

R: Reamplification



PiN outputs a low current

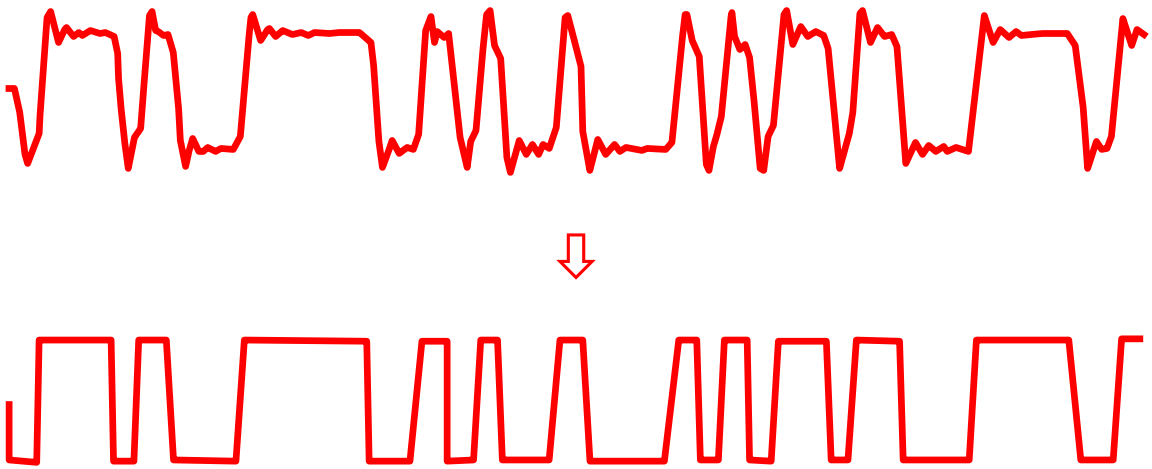


TIA Amplifies and converts to Voltage

2R: Reshaping

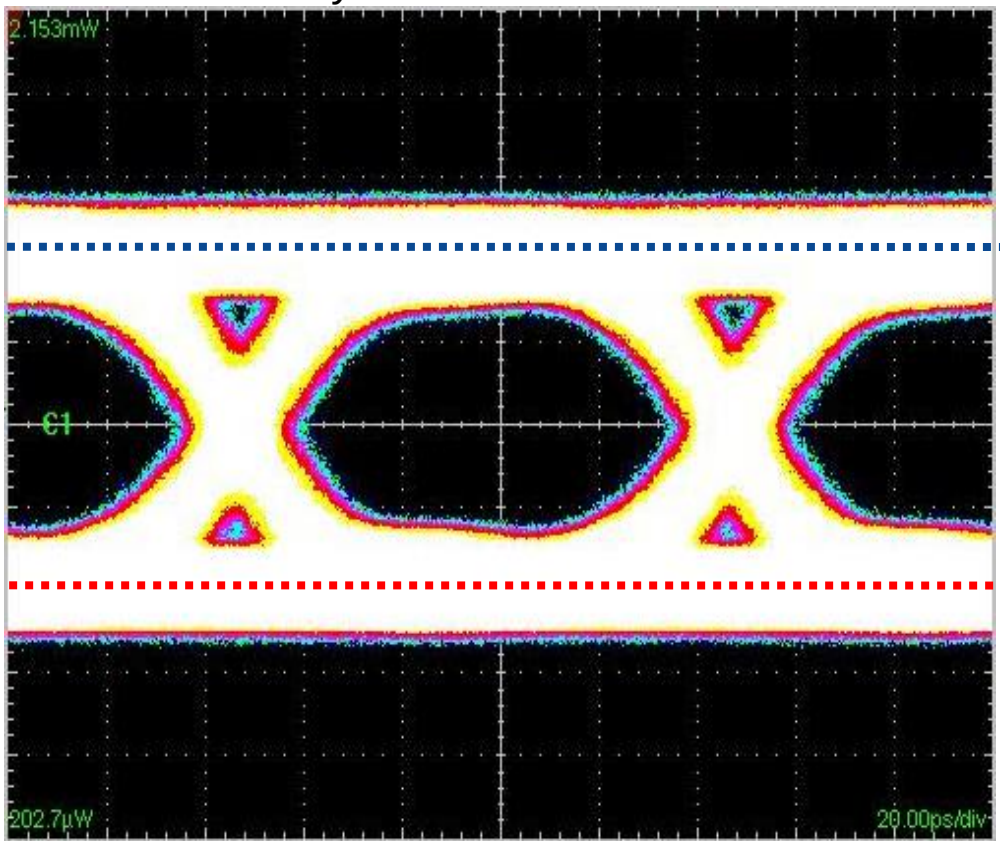
TIA is a linear amplifier

- If the input is above the decision threshold then output = 1
- If the input is below the decision threshold then output = 0



Noise

- The recovered signal has noise superimposed on it
  - Normally Gaussian Noise



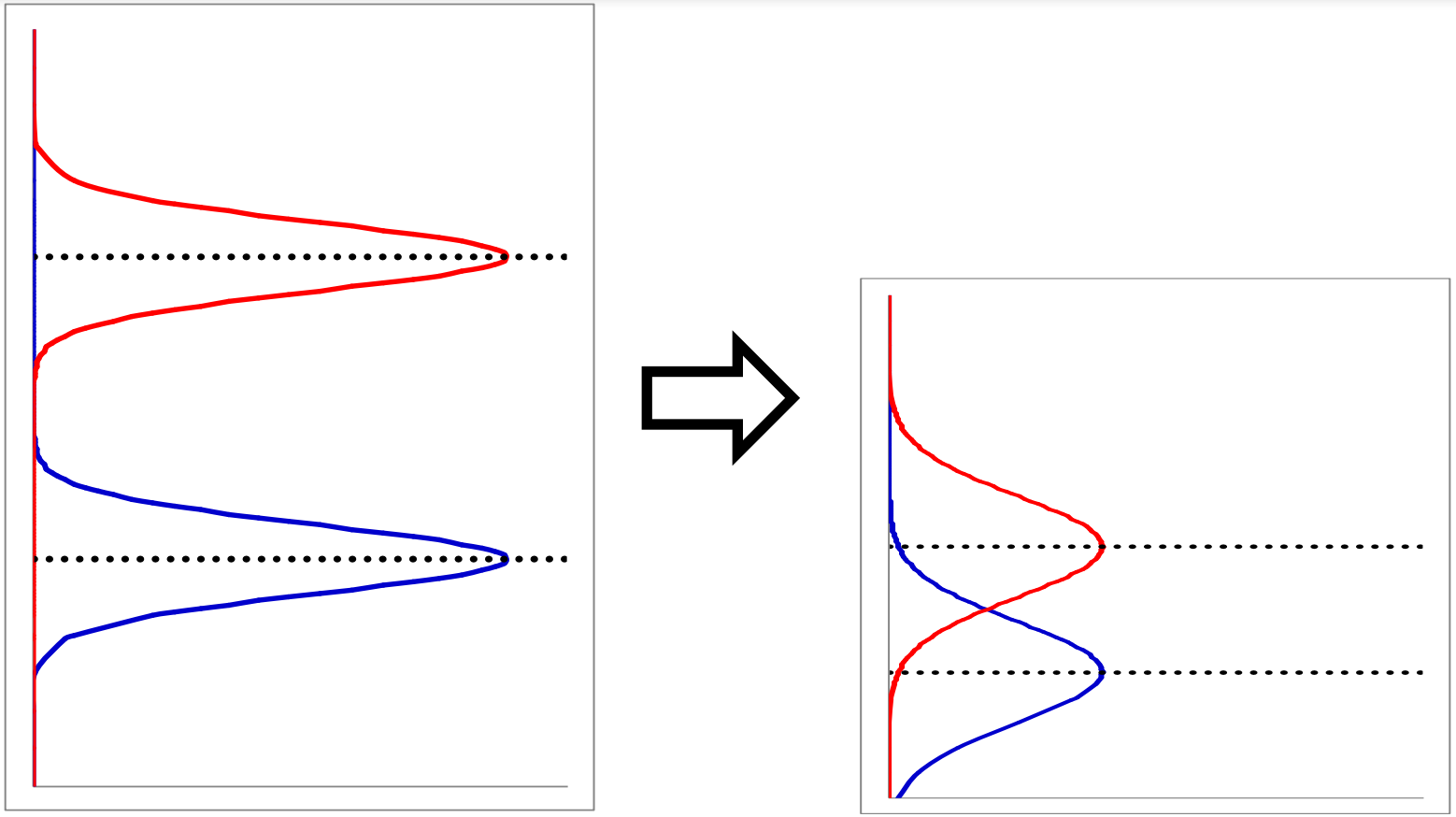
Noise on
1 level

Average Power of
1 level

Average Power of
0 level

Noise on
0 level

Signal to Noise Ratio



- As the power going in to the module decreases, the signal gets smaller, yet the noise essentially stays the same
  - The probability of getting an error increases

Errors = Latency

PCIe provides robust transmission

- **Guaranteed data transmission**
- **If there is an error, need to retransmit**

Obviously, causes latency but how much?

Latency Components

Latency Components

Physical Layer

Physics – Propagation Velocity

- **Velocity or Propagation (V_p) of a wave in a medium is given by:**

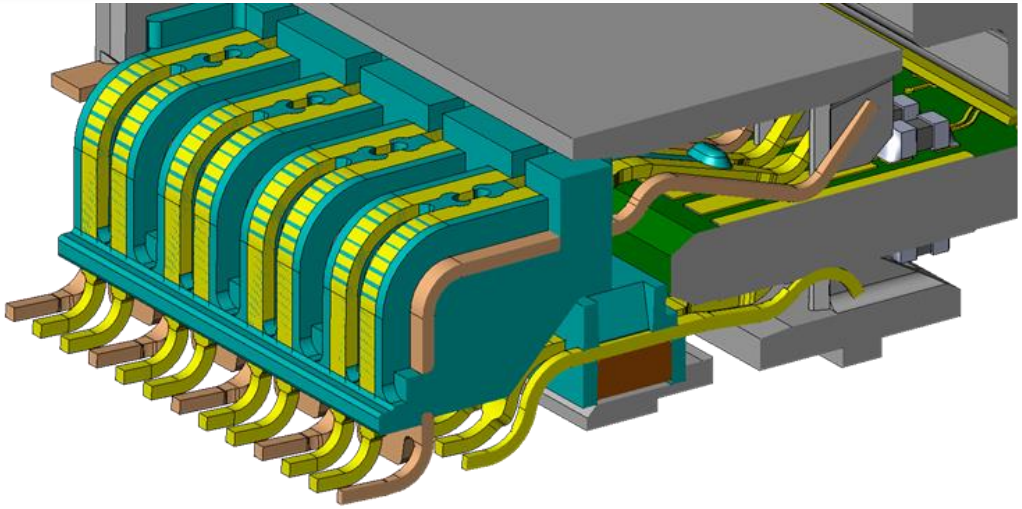
$$V_p = \frac{c}{\sqrt{\kappa}}$$

Where:

c = speed of light in a vacuum (299,792,458 m/s)

κ = dielectric constant of the medium

Connector System



Connector has a short row and a long row

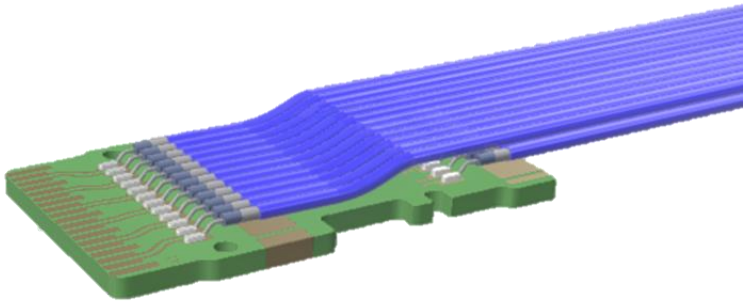
Full channel uses both, with one on each end

- Long row = 7.5 mm
- Short row = 5.7 mm

Propagation speed = 6.3 ns/m

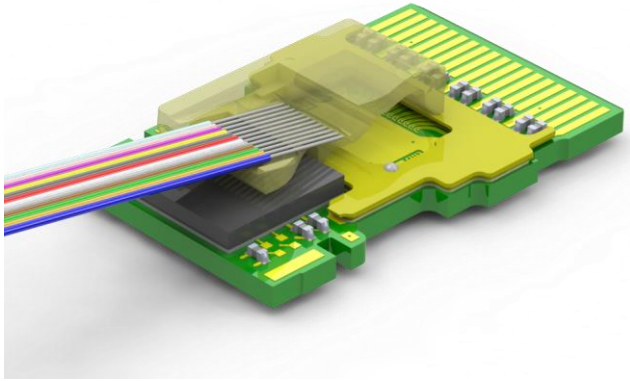
Total latency = 83 ps

Copper PCB

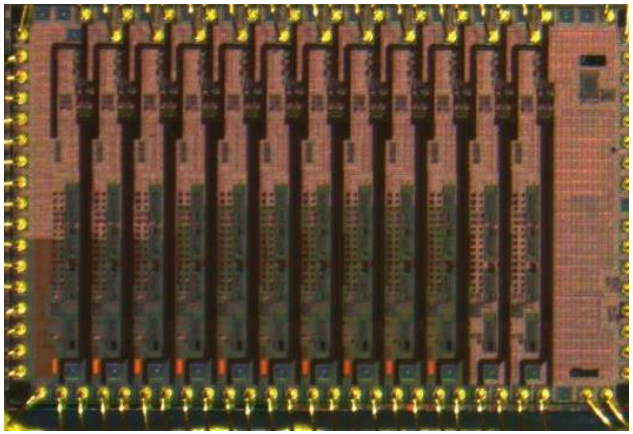


- **PCB propagation speed = 6.24 ns/m**
  - Short channel = 5.2 mm
  - Long channel = 11.2 mm
- **Again, full link uses both long and short**
- **Total latency = 102 ps**

Optical PCB



- **PCB propagation speed = 6.24 ns/m**
  - length = 6.36 mm
- **PCB latency = 61 ps**
- **ICs**
  - Feature and vendor specific
  - 100 – 500 ps per chip



Copper / Fiber

Nothing is faster than the speed of light in a vacuum

- Propagation speed in fiber is 5.13 ns/m
- Propagation speed in copper is 4.79 ns/m

Latency is dominated by length, *l*

- Copper = 0.18 + 4.8*l*
- Optical = 1.1 + 5.1*l*

Optical can do 100 m at 8 GT/s…

- Significant increase to “time of flight”

Latency Components

Link Layer
Transaction Layer

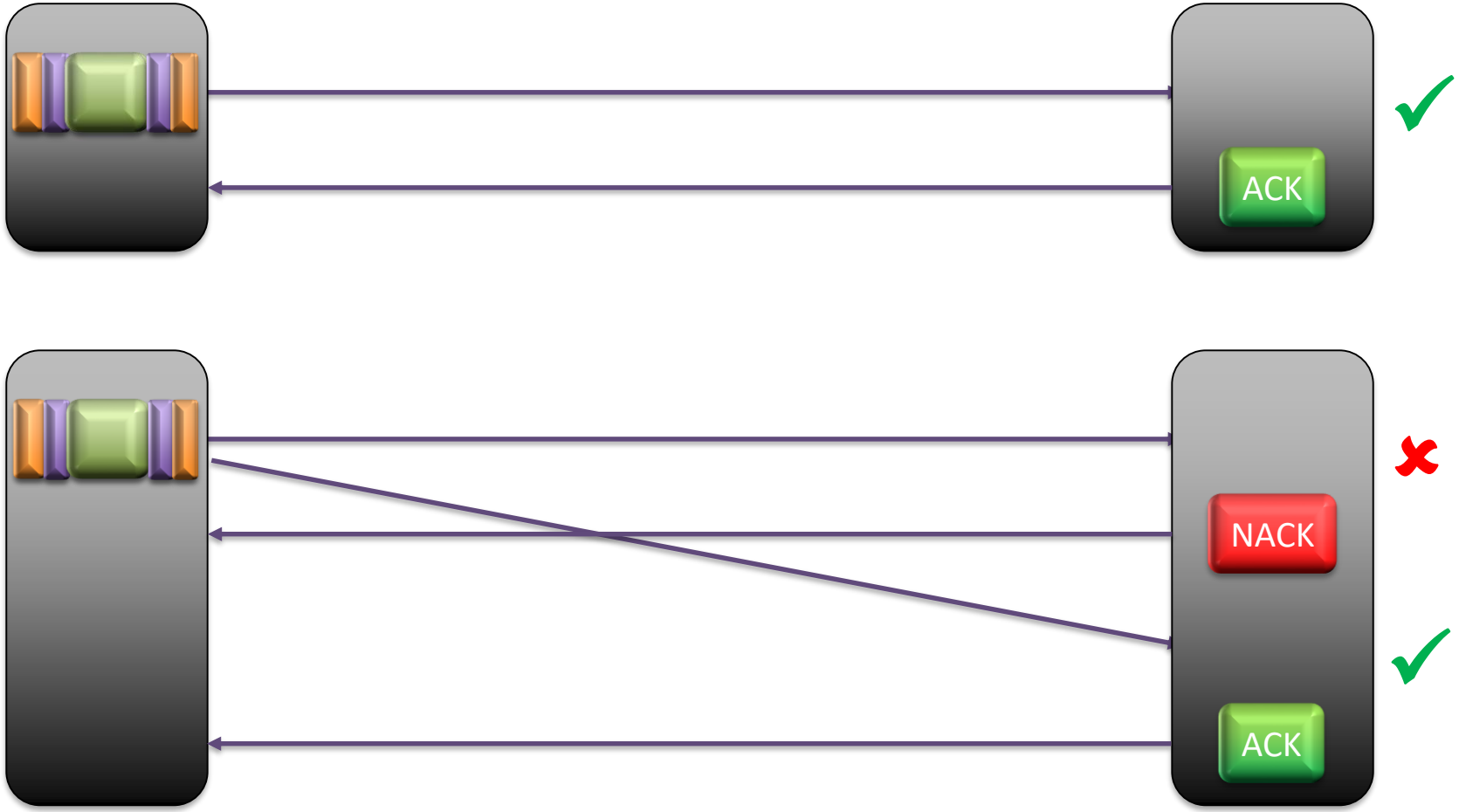
Robust Transmission

Key feature of PCIe is the inherent robustness of the transmission

- Sequence Number
- CRC

- ACK / NACK
- Flow Control

ACK / NACK



Sequence Flow

Latency for the NACKed data is > 3x that for a good packet

However packet ordering is also preserved

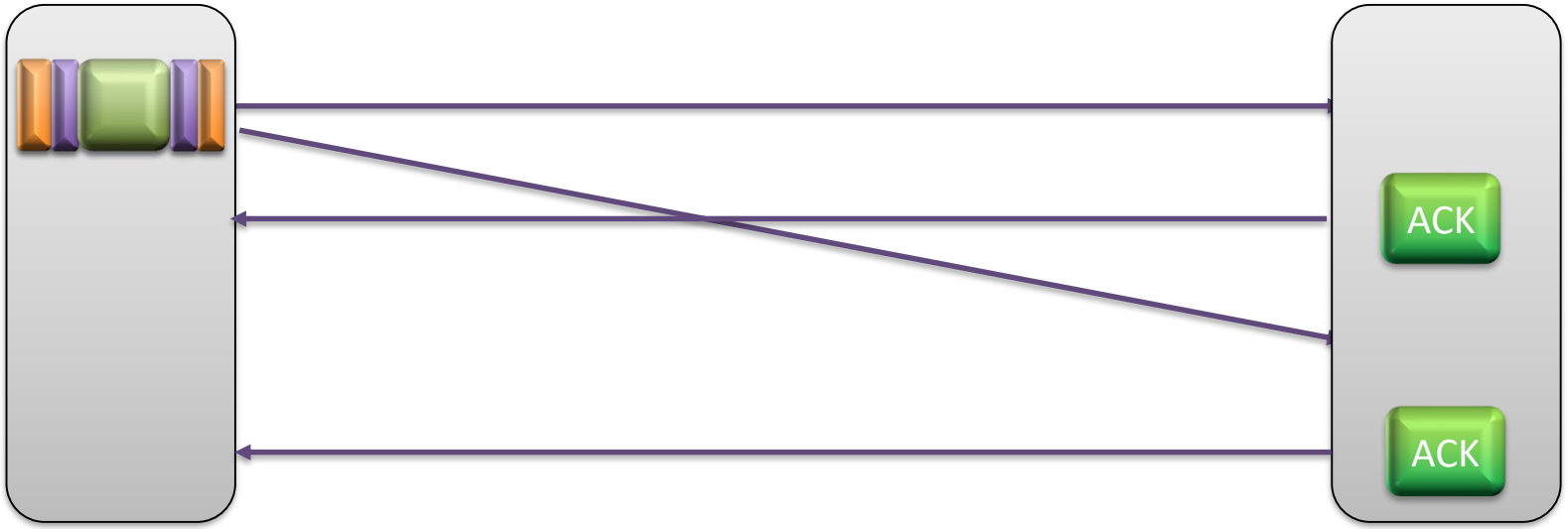
All subsequent TLP are discarded until the replayed DLPs are detected

ACK / NACK Replay Timer

Transmitter cannot assume a transaction has been correctly received

- Has to have an ACK

If an ACK is not received in an appropriate amount of time, the transmitter has to resend all unacknowledged TLPs



Buffer Size and Flow Control

PCIe control devices have to balance buffer size and cost

- **Too little results in increased latency**
- **Adding transistors adds cost**

Flow Control prevents data loss through credits

Flow Control

During initialization, each side reports it's buffer size

The transmitter can only send enough data to fill the buffer

It then needs to wait until there is an update on how many have been processed

Only then can it send more data

Data become bursty and latency increases

Latency Components

Operating System

Operating System

An application is allocated to a CPU and will run uninterrupted until it completes its tasks

- **More often regular operating systems will re-schedule the application or interrupt is when the OS issues an internal inter-CPU interrupt**

This is hardly noticeable in regular applications, but in latency critical systems timing requirements may be violated

- **Solution is to use a Real-Time Operating System that provides CPU shielding.**
  - An example is VxWorks®
- **Prevents interrupts**

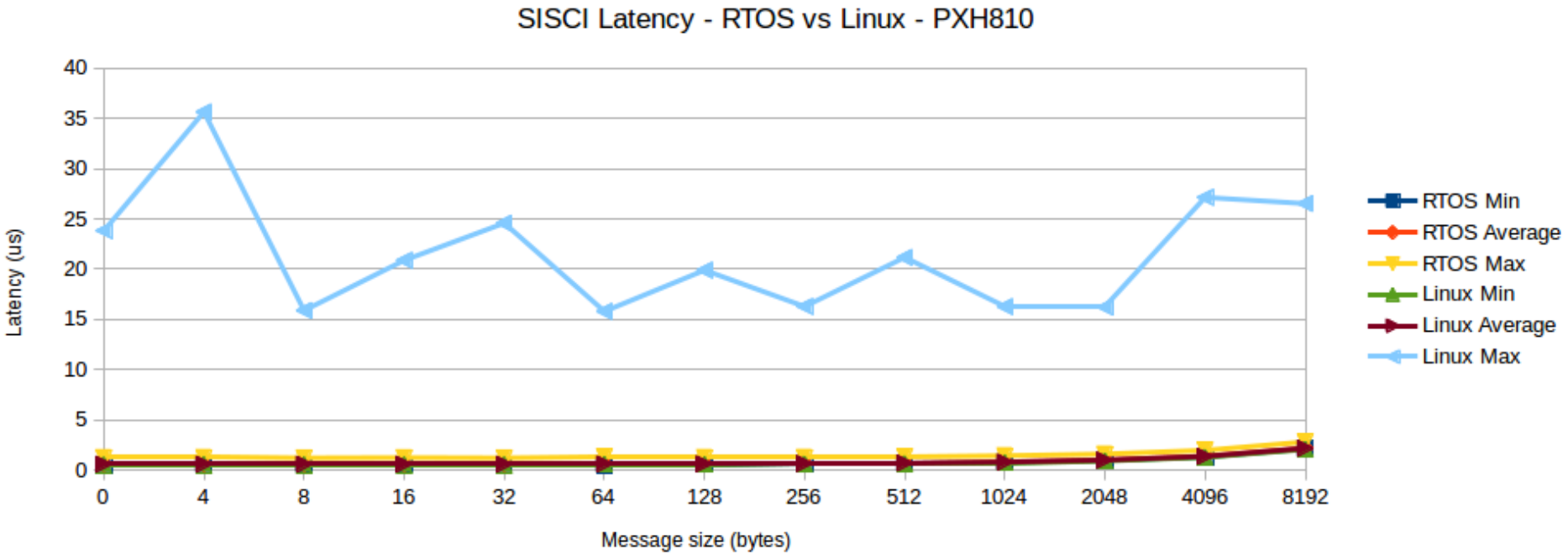
Measurements

Method

- **Used Dolphin's SISCI API ping pong benchmark to measure best, average and worst case latency for a sequence of ping pong data transfers**

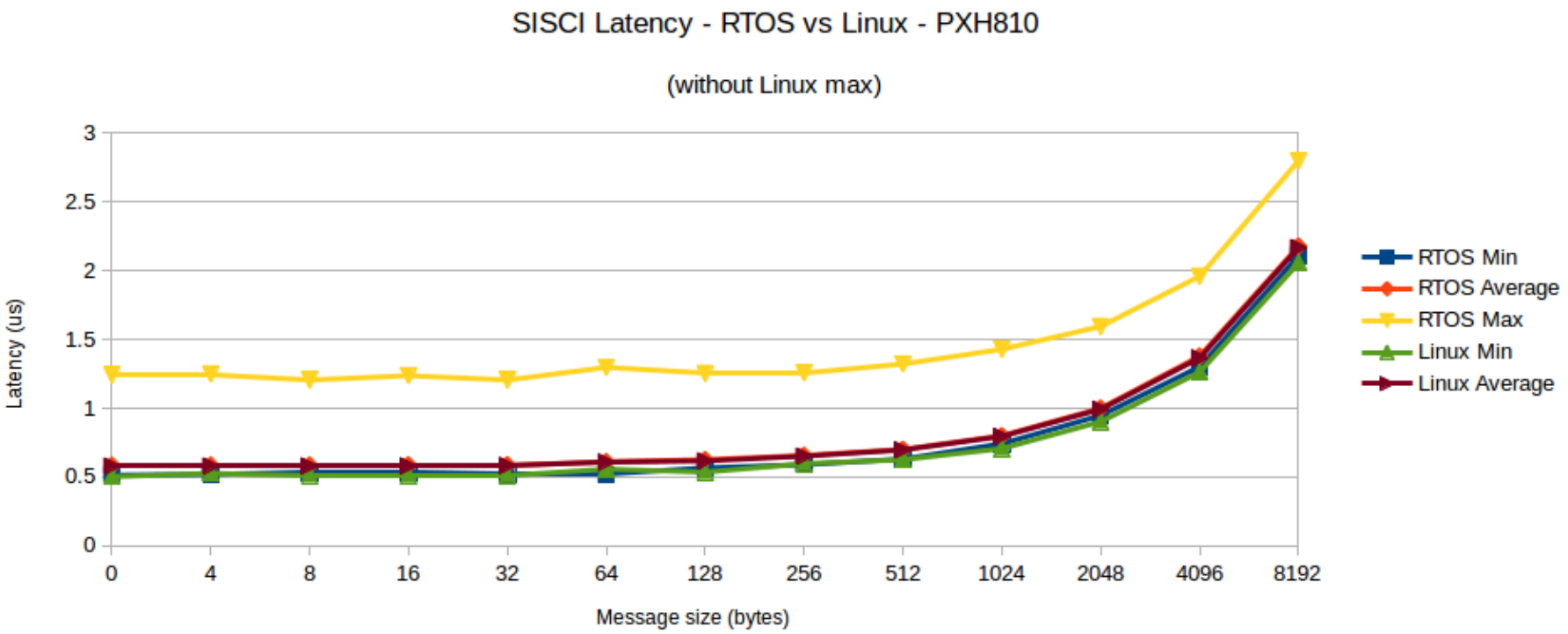
- **Latency is for a half-way round trip between two systems**
  - High precision timer used to measure the round-trip latency for each transfer

Comparing OS



- **Minimum = 510 ns**
- **Average = 540 ns**
- **Worst Case = 1.24 µs (VxWorks®)**
  35 µs (Linux)

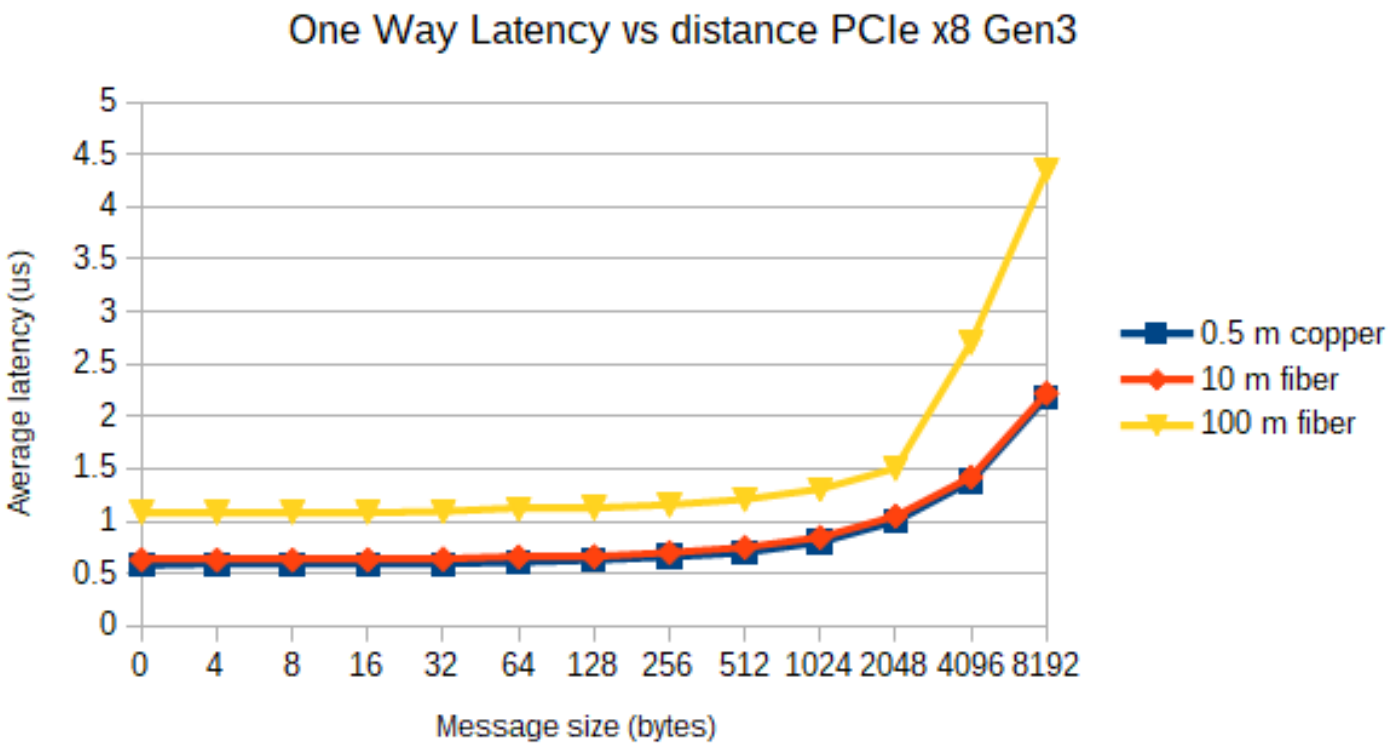
Same Results as Previous Without Linux



0.5 meter copper cable

VxWorks worst case latency is very close to average

Comparing Distance



4K and 8K message latency, 100 meter fiber

latency increases due to out of PCIe credits

Summary

Summary

Latency is a key benefit of PCIe systems

- **As links get longer time of flight becomes important**
  - Propagation delay
  - Protocol related resends and pauses
- **Copper / Optical latency generally follows the theory**
  - Latency disproportionally increases at long lengths and larger packet size due to flow control

Careful design of the interconnect, system and the channel can help minimise this

Thank you for attending the PCI-SIG Developers Conference 2018.

For more information, please go to www.pcisig.com

Don't forget to submit your feedback via the mobile app!

Download the **Crowd Compass** app and then search for **PCI-SIG Developers Conference** or entering the following URL into your mobile browser: https://crowd.cc/s/1rKy0
Enter event code: **DevCon2018**

Alternatively, access here: https://crowd.cc/pcisig2018

Note: Create an account within the app so Admin knows who to contact if selected as the prize winner.

Each session feedback is provided is equivalent to 1 raffle entry (up to 11 sessions).
General survey feedback = 1 raffle entry.